Statistics II - Final Data Analysis

Tarun Khanna 157762

1. Introduction and theory

This paper analyzes the effect of education on voter turnout. Education is a necessary tool that allows people to participate in the political process. It can be argued that education brings with it political awareness and encourages people to be more politically active. Thus education is expected to be positively related with the probability of a person voting. At an aggregate level as well, a more educated population should result in higher voter turnouts.

This theory is tested using panel data for 221 countries observed from 1945 to 2016. The ideal information to use would be the percentage of population with schooling or average years of schooling in each country. But in the given dataset the variables that contain such direct information on education are incomplete and result in too low degrees of freedom in the regression.

Therefore, the variable **percentage of total labor force with tertiary education (the highest level of educational attainment)** is used to operationalize the level of education. Another variable, 'educ', is created that captures the *percentage of total labor force that at least has secondary education*. It is generally expected that a more educated labor force would be associated with higher voter turnout.

We thus set out a regression with the null hypothesis that **Ho : percentage of labor force with education is not correlated with voter turnout**

2. Research Design

To estimate the effect of education on turnouts, a pooled OLS regression is carried out. But the estimates obtained using OLS are likely to be biased as a number of *unobserved factors* - cultural and institutional features of the country - also affect voter turnout. To control for these unobserved factors, a **fixed effects (FE)** regression for panel data is employed.

We also *control* for other factors that may vary with time and are therefore not controlled for by the FE estimation but are likely to be correlated with our independent variable and the dependent variable - **average GDP per capita (GDPPC)** and **unemployment rate**. GDPPC is expected to be negatively correlated with the dependent variable and positively with our key independent variable. On the other hand unemployment is expected to be negatively related to turnout, but negatively related to level of education. The equation that we estimte is thus equivalent to:

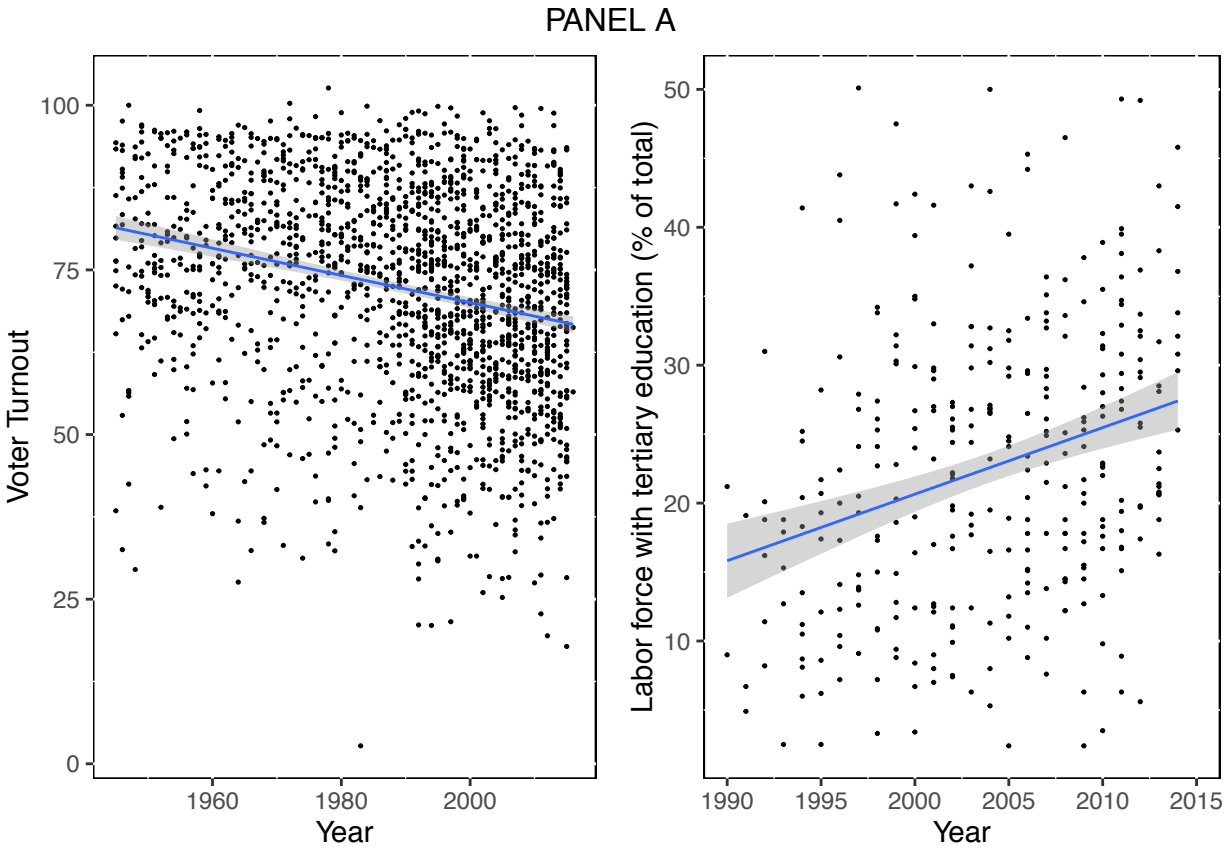
$$\alpha + \beta 1 * Education + \beta 2 * GDPPC + \beta 3 * UnemploymentRate : fixedeffects$$

A scatter plot of the dependent variable (Turnout)and the main independent variable (Percentage of labour force with Tertiary Education) across years is shown in *Panel A*. The two variables seem to be trending over time, which could lead to *spurious regression* results. In order to de-trend the series, another FE model specification with a **time trend** is also estimated. The equation that we estimte is thus equivalent to:

$$\alpha + \beta 1 * Education + \beta 2 * GDPPC + \beta 3 * UnemploymentRate + timetrend : fixedeffects$$

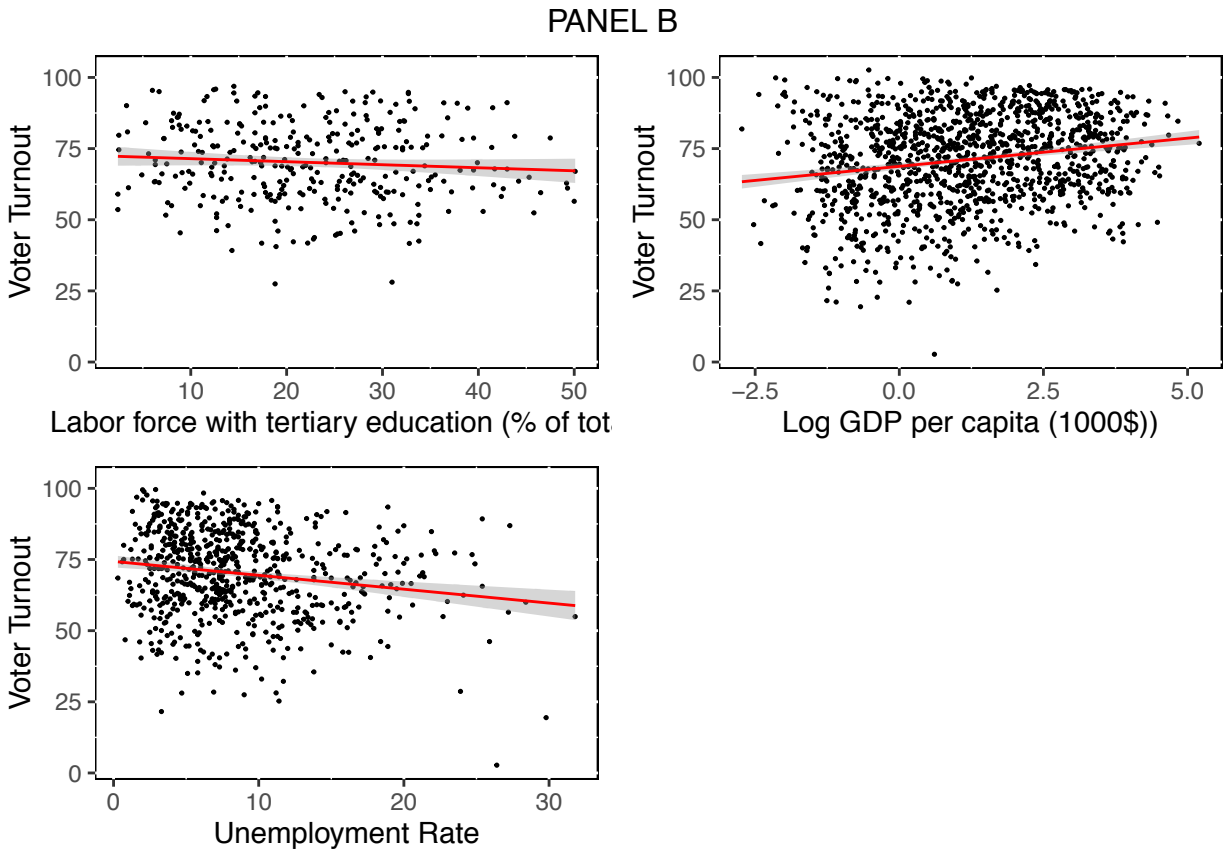
But the variation in turnout may not be uniform over time in panel units. Some years may have markedly higher or lower turnout. To control for such variation, time dummies for each year are added to the regression. The equation that we estimte is thus equivalent to:

$$\alpha + \beta 1 * Education + \beta 2 * GDPPC + \beta 3 * UnemploymentRate + \sum_{i=2}^{n} TimeDummies_i : fixedeffects$$



Panel B shows the plots between the independent variables and turnout. The distribution of GDP was right-skewed and the plot showed outlying observations with stong influence. This was corrected by taking the loged value of GDP is taken. The distribution of each variable seems random. Some of the observations for Unemployment Rate may be outliers.

Note: For the sake of completeness, a first difference and random effects model was also estimated (resuts not shown). A fixed effects model is preferred to a first difference model as we are dealing with an unbalanced panel, with high number of N and relatively small number of T. A fixed effects model is also preferred to a random effects model due to presence of 'entity fixed effects'. This is also confirmed by a Hausman Test.



3. Results

The table shows results of the regressions. Contrary to expectations, coefficient on education is *negative*. The coefficient on education indicates that an increase by one percentage point in the proportion of labor force with tertiary education *over time* is associated with decline in turnout by around 0.1 percentage points, controlling for differences across countries. But the value of the coefficient is *not statistically significant* at the usual levels of significance. This is true under all model specifications apart from the pooled OLS model (which is biased).**Overall it appears that education, at least higher education, is not a statistically significant predictor of turnout.**

With regards to the co-variates, the coefficient of loged GDPPC is negative and statistically significant at 5-10% level of significance depending on the model specification. The coefficient of unemployment rate is negative as expected but not statistically significant in all models apart from the model specification with time dummies.

The adjusted R2 is around 0.15 for the FE model. The model with time trend incorporated does not significantly change results, the time trend is also not statistically significant. The adjusted R2 increases significantly to 0.25 when time dummies are included even though none of the dummies are statistically significant (results supressed). The substantive and statistical significance of coefficients also increases in this specification.

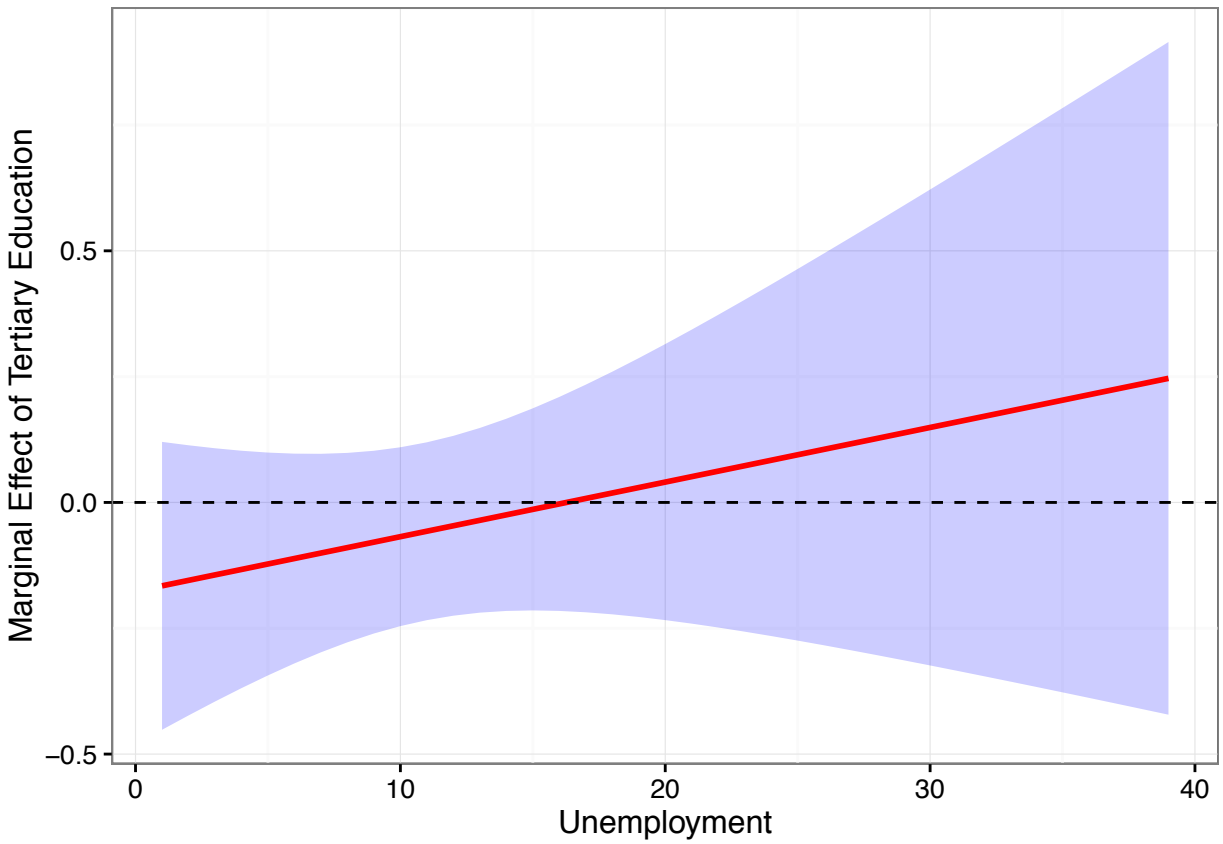
	Dependent variable:			
	Voter Turnout			
	(1)	(2)	(3)	(4)
Tertiary Education	-0.264^{***}	-0.092	-0.071	-0.118
	(0.083)	(0.080)	(0.092)	(0.095)
GDPPC (1000)	2.990^{***}	-4.674^{***}	-3.975^{**}	-6.750^{***}
	(0.721)	(0.970)	(1.790)	(2.124)
Unemployment	-0.278^{*}	-0.212	-0.197	-0.306^{*}
	(0.157)	(0.148)	(0.151)	(0.168)
Time			-0.070	
			(0.151)	
Observations	298	298	298	298
R^2	0.089	0.144	0.145	0.249
Adjusted R^2	0.088	0.100	0.100	0.154
F Statistic	9.633^{***}	11.559^{***}	8.691^{***}	2.438^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

(1) pooled OLS (2) Fixed effects (3) FE with time trend (4) FE with time dummies (suppressed)

Interactions

Theory suggests that the effect of tertiary education on turnout may be conditioned by unemployment. The effect may be stronger in countries with high level of unemployment. The model results (not shown here) with the interaction shows only a small positive marginal effect, which is also not statistically significant at the usual levels. *But the significance of the interaction term is not a reliable method to check for the marginal effect.* A plot of the marginal effect of education vs. the unemployment rate shows that zero is in the 95% confidence interval. This suggests that the marginal effect is not significant at any level of unemployment.



Testing for appropriateness of method

The following tests are carried out to establish the appropriateness of the techniques used:

- **Testing for Panel effects**: A Lagrange-Multiplier with the null hypothesis that there is no significant difference across units (i.e. no panel effect) was employed. The test returns a p value < 0.001, suggesting presence of both panel effects.
- **Testing for Random vs. Fixed effects**: A Hausman Test with the null hypothesis that the preferred model is random effects vs. the alternative the fixed effects was used. The test returns p value <0.001. Therefore, fixed effects is preferred over random effects.

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  vt ~ wdi_lfte + log(unna_gdppc.1) + wdi_unempne
## chisq = 1066500, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects

## series eltype, vi_rsg, vi_rsm are constants and have been removed

##
##  Hausman Test
##
## data:  vt ~ wdi_lfte + log(unna_gdppc.1) + wdi_unempne
## chisq = 46.808, df = 3, p-value = 3.819e-10
## alternative hypothesis: one model is inconsistent
```

3. Diagnostics

- **Multicollinearity**: A VIF test confirms that there is no significant collinearity between the explanatory variables.
- **Unit Root**: The Dickey-Fuller is used test to check for stochastic trends. The null hypothesis is that the series has a unit root (i.e. non-stationary). The test returns a p-value <0.01. We therefore reject the null and presence of unit roots.
- **Serial correlation**: Such tests apply usually to macro panels with long time series. It is generally not a problem in micro panels (with very few years). A Breusch-Godfrey/Wooldridge test for serial correlation in panel models is applied. The null is that there is no serial correlation is rejected at 5 % level of significance. The test returns a p value > 0.01 for the model with time trend, indicating no serial correlation. But the model with time dummes returns a p value <0.001 indicating serial correlation. This might decrease the standard errors and may explain the high statistical significance of the variables in the regression results.
- **Heteroskedasticity**: A Breusch-Pagan test is carried out with the null hypothesis being homoskedasticity. A p-value < 0.01 indicated presence of Heteroskedasticity.

```
##          wdi_lfte log(unna_gdppc.1)       wdi_unempne
##          1.253289          1.385623          1.123496
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  vt ~ wdi_lfte + log(unna_gdppc.1) + wdi_unempne + factor(year)
## chisq = 12.213, df = 1, p-value = 0.0004747
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
## series    are constants and have been removed
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  Panel.set$vt
## Dickey-Fuller = -13.09, Lag order = 2, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  Breusch-Pagan test
##
## data:  vt ~ wdi_lfpe + factor(country)
## BP = 405.28, df = 192, p-value < 2.2e-16
```

In order to correct for serial correlation in errors and heteroskedasticity we estimate robust standard errors (using the vcovHC "arellano" estimator which corrects for both heteroskedasticity and serial correlation and is recommended for fixed effects). There is however no significant variation in regression results and levels of significance of the variables.

4. Robustness

To test the robustness of the results the main independent variable is replaced with another approximate measure of education - 'educ' which captures the percentage of labour force with at least a secondary level of education.

An extended model is also estimated by including two new independent variables - percentage of rural population (which is expected to be negatively related to turnout) and public health expenditure (which is expected to be positively related to turnout).

The model with the new dependant variable gives similar results to the previously estimated model. The coefficient of education is still not significant. Adding the variables, increases the coefficient of education but it is still not significant. **As such the results seem robust to change in dependent variable estimator or inclusion of new covariates.**

	Dependent variable:				
	Voter Turnout				
	(1)	(2)	(3)	(4)	(5)
Tertiary Education	−0.092	−0.071	−0.118		−0.127
	(0.080)	(0.092)	(0.095)		(0.112)
At least Secondary Educ				−0.117	
				(0.078)	
GDPPC (1000)	−4.674***	−3.975**	−6.750***	−5.674***	−5.763**
	(0.970)	(1.790)	(2.124)	(2.092)	(2.471)
Unemployment	−0.212	−0.197	−0.306*	−0.287*	−0.310
	(0.148)	(0.151)	(0.168)	(0.167)	(0.188)
Time		−0.070			
		(0.151)			
Observations	298	298	298	295	269
R^2	0.144	0.145	0.249	0.257	0.200
Adjusted R^2	0.100	0.100	0.154	0.159	0.119
F Statistic	11.559***	8.691***	2.438***	2.522***	1.823**

Note: *p<0.1; **p<0.05; ***p<0.01

(1) FE (2) FE w/ time trend (3) FE w/ time dummies (4) Educ as DV (5) Extended Model

5. Conclusion

It was expected that higher percentage of labor force with education would be associated with higher voter turnout. This is not supported by the panel data available. We cannot reject the null-hypothesis.

At best a **marginally negative effect, which is not significant statistically**, is found. Can the negative relationship be explained? Some studies find that as the level of educational achievement increases beyond regular schooling, the propensity of individual to vote declines. This might indicate the apathy of the educated elite or their disenchantment with the political process. Given that the main explanatory variable of this analysis was percentage of population with tertiary education, *which captures college level education*, the results are not surprising and seem to fit with earlier findings in this regard.

```
---
title: "Statistics II - Final Data Analysis"
author: "Tarun Khanna 157762"
output:
  pdf_document:
    keep_tex: yes
  html_document: default
  word_document: default
---

```{r options, include=T, message=F, echo=FALSE, warning=FALSE, error=FALSE}

knitr::opts_chunk$set(echo = TRUE, eval = TRUE, message = FALSE)
library(foreign)  # for using read.dta() to open the .dta-file
library(stargazer)  # for summary statistics and regression tables
library(magrittr)  # for 'piping': more readable code
library(ggplot2)  # the ggplot2 package provides nice function for plotting
#library(dplyr)  # for data manipulation
library(car) #for recode
#library(sandwich) #for robust standard errors
#library(lmtest) #for robust standard errors
library(plm) #for panel
library(gridExtra)

```

1. Introduction and theory

This paper analyzes the effect of education on voter turnout. Education is a necessary tool that allows people to participate in the political process. It can be argued that education brings with it political awareness and encourages people to be more politically active. Thus education is expected to be positively related with the probability of a person voting. At an aggregate level as well, a more educated population should result in higher voter turnouts.

This theory is tested using panel data for 221 countries observed from 1945 to 2016. The ideal information to use would be the percentage of population with schooling or average years of schooling in each country. But in the given dataset the variables that contain such direct information on education are incomplete and result in too low degrees of freedom in the regression.

Therefore, the variable __percentage of total labor force with tertiary education (the highest level of educational attainment)__ is used to operationalize the level of education. Another variable, 'educ', is created that captures the _percentage of total labor force that at least has secondary education_. It is generally expected that a more educated labor force would be associated with higher voter turnout.

We thus set out a regression with the null hypothesis that __Ho : percentage of labor force with education is not correlated with voter turnout__

2. Research Design
```

To estimate the effect of education on turnouts, a pooled OLS regression is carried out. But the estimates obtained using OLS are likely to be biased as a number of _unobserved factors_ - cultural and institutional features of the country - also affect voter turnout. To control for these unobserved factors, a __fixed effects (FE)__ regression for panel data is employed.

We also _control_ for other factors that may vary with time and are therefore not controlled for by the FE estimation but are likely to be correlated with our independent variable and the dependent variable - __average GDP per capita (GDPPC)__ and __unemployment rate__. GDPPC is expected to be negatively correlated with the dependent variable and positively with our key independent variable. On the other hand unemployment is expected to be negatively related to turnout, but negatively related to level of education. The equation that we estimte is thus equivalent to: $$ \alpha + \beta1*Education + \beta2*GDPPC + \beta3*Unemployment Rate : fixed effects $$

A scatter plot of the dependent variable (Turnout)and the main independent variable (Percentage of labour force with Tertiary Education) across years is shown in _Panel A_. The two variables seem to be trending over time, which could lead to _spurious regression_ results. In order to de-trend the series, another FE model specification with a __time trend__ is also estimated. The equation that we estimte is thus equivalent to: $$ \alpha + \beta1*Education + \beta2*GDPPC + \beta3*Unemployment Rate + time trend : fixed effects $$

##
##
##

But the variation in turnout may not be uniform over time in panel units. Some years may have markedly higher or lower turnout. To control for such variation, time dummies for each year are added to the regression. The equation that we estimte is thus equivalent to: $$ \alpha + \beta1*Education + \beta2*GDPPC + \beta3*Unemployment Rate + \sum_{i=2}^n Time Dummies_i : fixed effects $$

```{r include=T, message=F, echo=FALSE, warning=FALSE, error=FALSE}

setwd('/Users/tarunkhanna/Dropbox/FDA')
d <- read.dta('panel.dta', convert.factors = F)

# this gives a table of variable names and labels
dataKey <- data.frame(varName=names(d), varLabel=attr(d, "var.labels"))

# this gives you a list containing all value labels
valueLabels <- attr(d, 'label.table')

# subsetting data to remove missing for dependent variables
d <- d[-which(is.na(d$vt)),  ]
d$educ = d$wdi_lfse + d$wdi_lfte
d$unna_gdppc.1 = d$unna_gdppc/1000

# pre-regression plots
```

```
p1 <- ggplot(d, aes(x=year, y=vt), width=1, height=1)+geom_point(size
=0.25)+geom_smooth(method="lm",  size =.5)+xlab('Year') + ylab('Voter
Turnout')+theme(panel.background = element_rect(fill = 'white', colour =
'black'))

p2 <- ggplot(d, aes(x=year, y=wdi_lfte))+geom_point(size
=0.25)+geom_smooth(method="lm", size =.5) + xlim(1990, 2015)+ylab('Labor
force with tertiary education (% of total)') +
xlab('Year')+theme(panel.background = element_rect(fill = 'white', colour =
'black'))

grid.arrange(p1,p2, ncol=2, top = "PANEL A")

```

Panel B shows the plots between the independent variables and turnout. The distribution of GDP was right-skewed and the plot showed outlying observations with stong influence. This was corrected by taking the loged value of GDP is taken. The distribution of each variable seems random. Some of the observations for Unemployment Rate may be outliers.

__Note__: For the sake of completeness, a first difference and random effects model was also estimated (resuts not shown). A fixed effects model is preferred to a first difference model as we are dealing with an unbalanced panel, with high number of N and relatively small number of T. A fixed effects model is also preferred to a random effects model due to presence of 'entity fixed effects'. This is also confirmed by a Hausman Test.

```{r include=T, message=F, echo=FALSE, warning=FALSE, error=FALSE}

p3 <- ggplot(data =d, mapping = aes(wdi_lfte, vt)) + geom_point(size = 0.25)
+ geom_smooth(method="lm",colour = "red",size =.5) + xlab('Labor force with
tertiary education (% of total)') + ylab('Voter
Turnout')+theme(panel.background = element_rect(fill = 'white', colour =
'black'))

p4 <- ggplot(data =d, mapping = aes(log(unna_gdppc.1), vt)) + geom_point(size
= 0.25) + geom_smooth(method="lm",colour = "red",size =.5) + xlab('Log GDP
per capita (1000$))') + ylab('Voter Turnout')+theme(panel.background =
element_rect(fill = 'white', colour = 'black'))

p5 <- ggplot(data =d, mapping = aes(wdi_unempne, vt)) + geom_point(size =
0.25) + geom_smooth(method="lm",colour = "red",size =.5) + xlab('Unemployment
Rate') + ylab('Voter Turnout')+theme(panel.background = element_rect(fill =
'white', colour = 'black'))

#ggplot(d) + geom_bar(aes(year, vt, fill = as.factor(year)), position =
"dodge", stat = "summary", fun.y = "mean")

grid.arrange(p3,p4,p5, ncol=2, top = "PANEL B")

```

3. Results

The table shows results of the regressions. Contrary to expectations, coefficient on education is _negative_. The coefficient on education indicates that an increase by one percentage point in the proportion of labor force with tertiary education _over time_ is associated with decline in turnout by around 0.1 percentage points, controlling for differences across countries. But the value of the coefficient is _not statistically significant_ at the usual levels of significance. This is true under all model specifications apart from the pooled OLS model (which is biased).__Overall it appears that education, at least higher education, is not a statistically significant predictor of turnout.__

With regards to the co-variates, the coefficient of loged GDPPC is negative and statistically significant at 5-10% level of significance depending on the model specification. The coefficient of unemployment rate is negative as expected but not statistically significant in all models apart from the model specification with time dummies.

The adjusted R2 is around 0.15 for the FE model. The model with time trend incorporated does not significantly change results, the time trend is also not statistically significant. The adjusted R2 increases significantly to 0.25 when time dummies are included even though none of the dummies are statistically significant (results supressed). The substantive and statistical significance of coefficients also increases in this specification.

```{r include=T, message=F, echo=FALSE, warning=FALSE, error=FALSE, results="hide"}

m.pols <- plm(vt ~ wdi_lfte  + log(unna_gdppc.1) + wdi_unempne , data=d, index=c("country", "year"), model="pooling")

m.fe <- plm(vt ~ wdi_lfte  + log(unna_gdppc.1) + wdi_unempne, data=d, index=c("country", "year"), model="within")
summary(fixef(m.fe)) # to see entity dummy values

d$time<- recode (d$year, "1990=1;
1991=2;1992=3;1993=4;1994=5;1995=6;1996=7;1997=8;
1998=9;1999=10;2000=11;2001=12;2002=13;2003=14;2004=15;2005=16;2006=17;2007=18;

2008=19;2009=20;2010=21;2011=22;2012=23;2013=24;2014=25;") # adding a time trend variable

m.fe.trend <- plm(vt ~ wdi_lfte  + log(unna_gdppc.1) + wdi_unempne + time , data=d, index=c("country", "year"), model="within")

m.fe.time <- plm(vt ~ wdi_lfte  + log(unna_gdppc.1) + wdi_unempne + factor(year) , data=d, index=c("country", "year"), model="within")

#m.fe.two <- plm(vt ~ wdi_lfte + log(unna_gdppc.1) + wdi_unempne , data=d, index=c("country", "year"), model="within", effect = 'twoways')
#summary(fixef(m.fe.two,type="dfirst",effect="time")) # to see time dummy values
#summary(fixef(m.fe.two,type="dfirst",effect="individual")) # to see entity
```

dummy values

```

```{r  include = T, echo=FALSE, results='asis'}

stargazer(m.pols, m.fe, m.fe.trend, m.fe.time, df = FALSE, keep =
c('wdi_lfte', 'unna_gdppc.1', 'wdi_unempne', 'time' ), dep.var.labels =
'Voter Turnout', covariate.labels = c('Tertiary Education', 'GDPPC (1000)',
'Unemployment','Time'), notes = '(1) pooled OLS (2) Fixed effects (3) FE with
time trend (4) FE with time dummies (suppressed)', header = F, float = F, out
= "reg.html")

```

Interactions

Theory suggests that the effect of tertiary education on turnout may be conditioned by unemployment. The effect may be stronger in countries with high level of unemployment. The model results (not shown here) with the interaction shows only a small positive marginal effect, which is also not statistically significant at the usual levels. _But the significance of the interaction term is not a reliable method to check for the marginal effect._ A plot of the marginal effect of education vs. the unemployment rate shows that zero is in the 95% confidence interval. This suggests that the marginal effect is not significant at any level of unemployment.

```{r , include=T, message=F, warning=FALSE, error=FALSE, echo=FALSE,
results= "hide"}

m.fe.i <- plm(vt ~ wdi_lfte*wdi_unempne + unna_gdppc.1 + time, data=d,
index=c("country", "year"), model="within")

vb1 <- m.fe.i %>% vcov %>% diag %>% .[1]
vb3 <- m.fe.i %>% vcov %>% diag %>% .[5]
cvb1b3 <- m.fe.i %>% vcov %>% .[1,5]
margins <- data.frame(
  unem = 1:max(d$wdi_unempfne, na.rm =T),
  me = coef(m.fe.i)[1] + coef(m.fe.i)[5] * (1:max(d$wdi_unempfne, na.rm =T)),
  se = sqrt(vb1 + (1:max(d$wdi_unempfne, na.rm =T))^2 * vb3 + 2 *
(1:max(d$wdi_unempfne, na.rm =T)) * cvb1b3)
)
margins$lwr <- margins$me - 1.96 * margins$se
margins$upr <- margins$me + 1.96 * margins$se

ggplot(data = margins, aes(x = unem, y = me, ymin = lwr, ymax = upr)) +
  geom_ribbon(fill = 'blue', alpha = 0.2) + geom_line(colour = "red", size
=1) + theme_bw() + geom_hline(yintercept = 0, linetype = 'dashed') +
xlab('Unemployment') + ylab('Marginal Effect of Tertiary
Education')+theme(panel.background = element_rect(fill = 'white', colour =
'black'))

```
```

Testing for appropriateness of method

The following tests are carried out to establish the appropriateness of the techniques used:

* __Testing for Panel effects__: A Lagrange-Multiplier with the null hypothesis that there is no significant difference across units (i.e. no panel effect) was employed. The test returns a p value < 0.001, suggesting presence of both panel effects.

* __Testing for Random vs. Fixed effects__: A Hausman Test with the null hypothesis that the preferred model is random effects vs. the alternative the fixed effects was used. The test returns p value <0.001. Therefore, fixed effects is preferred over random effects.

```{r , include=T, message=F, warning=FALSE, error=FALSE, echo=FALSE}

plmtest(m.pols, type=("bp")) # to test for panel effect

m.re <- plm(vt ~ wdi_lfte + log(unna_gdppc.1) + wdi_unempne, data=d,
index=c("country", "year"), model="random")
phtest(m.fe, m.re) # Hausman test for fixed effects

```

3. Diagnostics

* __Multicollinearity__: A VIF test confirms that there is no significant collinearity between the explanatory variables.

* __Unit Root__: The Dickey-Fuller is used test to check for stochastic trends. The null hypothesis is that the series has a unit root (i.e. non-stationary). The test returns a p-value <0.01. We therefore reject the null and presence of unit roots.

* __Serial correlation__: Such tests apply usually to macro panels with long time series. It is generally not a problem in micro panels (with very few years). A Breusch-Godfrey/Wooldridge test for serial correlation in panel models is applied. The null is that there is no serial correlation is rejected at 5 % level of significance. The test returns a p value > 0.01 for the model with time trend, indicating no serial correlation. But the model with time dummes returns a p value <0.001 indicating serial correlation. This might decrease the standard errors and may explain the high statistical significance of the variables in the regression results.

* __Heteroskedasticity__: A Breusch-Pagan test is carried out with the null hypothesis being homoskedasticity. A p-value < 0.01 indicated presence of Heteroskedasticity.

```{r include=T, message=F, echo=FALSE, warning=FALSE, error=FALSE}

vif(m.pols) # testing for collinearity

pbgtest(m.fe.time) # testing formserial correlation
```

```
library(tseries) # The Dickey-Fuller for unit roots
Panel.set <- plm.data(d, index = c("country", "year")) # The Dickey-Fuller
for unit roots
adf.test(Panel.set$vt, k=2) # The Dickey-Fuller for unit roots

library(lmtest) #Breusch-Pagan test for heteroskedasticity
bptest(vt ~ wdi_lfpe + factor(country), data = d, studentize=F) #Breusch-
Pagan test for heteroskedasticity

c1 <- coeftest(m.fe.time)      # Original coefficients
c2 <- coeftest(m.fe.time, vcovHC) # Heteroskedasticity consistent
coefficients
c3 <- coeftest(m.fe.time, vcovHC(m.fe.time, method = "arellano")) #
Heteroskedasticity consistent coefficients (Arellano)
```

In order to correct for serial correlation in errors and heteroskedasticity we estimate robust standard errors (using the vcovHC "arellano" estimator which corrects for both heteroskedasticity and serial correlation and is recommended for fixed effects). There is however no significant variation in regression results and levels of significance of the variables.

4. Robustness

To test the robustness of the results the main independent variable is replaced with another approximate measure of education - 'educ' which captures the percentage of labour force with at least a secondary level of education.

An extended model is also estimated by including two new independent variables - percentage of rural population (which is expected to be negatively related to turnout) and public health expenditure (which is expected to be positively related to turnout).

The model with the new dependant variable gives similar results to the previously estimated model. The coefficient of education is still not significant. Adding the variables, increases the coefficient of education but it is still not significant. __As such the results seem robust to change in dependent variable estimator or inclusion of new covariates.__

```{r , include=T, message=F, echo=FALSE, warning=FALSE, error=FALSE,
results= "hide"}

mfe_robust.t <- plm(vt ~ educ + log(unna_gdppc.1) + wdi_unempne +
factor(year) , data=d, index=c("country", "year"), model="within")

mfe_robust.t2 <- plm(vt ~ wdi_lfte + log(unna_gdppc.1) + wdi_unempne +
factor(year) + une_rp + wdi_hepub , data=d, index=c("country", "year"),
model="within")

```

```{r , include=T, message=F, echo=FALSE, results='asis'}
```

```
stargazer(m.fe, m.fe.trend, m.fe.time, mfe_robust.t, mfe_robust.t2, df =
FALSE, keep = c('wdi_lfte','educ', 'unna_gdppc.1','wdi_unempne', 'time' ),
dep.var.labels = 'Voter Turnout', covariate.labels = c('Tertiary Education',
'At least Secondary Educ','GDPPC (1000)', 'Unemployment','Time','Rural
population', 'Health Exp'), notes = '(1) FE (2) FE w/ time trend (3) FE w/
time dummies (4) Educ as DV (5) Extended Model', header = F, float = F, out =
"reg.html")
```

5. Conclusion

It was expected that higher percentage of labor force with education would be associated with higher voter turnout. This is not supported by the panel data available. We cannot reject the null-hypothesis.

At best a __marginally negative effect, which is not significant statistically__, is found. Can the negative relationship be explained? Some studies find that as the level of educational achievement increases beyond regular schooling, the propensity of individual to vote declines. This might indicate the apathy of the educated elite or their disenchantment with the political process. Given that the main explanatory variable of this analysis was percentage of population with tertiary education, _which captures college level education_, the results are not surprising and seem to fit with earlier findings in this regard.